Classification vs Regression

Classification vs Regression

Which type of supervised machine learning problem is this, Classification or regression? Why?

This is a classification problem because we are trying to predict if student will drop out or not based on aviable data

Exploring the Data

Data exploration

Student response addresses the most important characteristics of the dataset and uses these characteristics to inform their decision making. Important characteristics must include:

```
Total number of students: 395
Number of students who passed: 265
Number of students who failed: 130
Number of features: 30
Graduation rate of the class: 67.09%
```

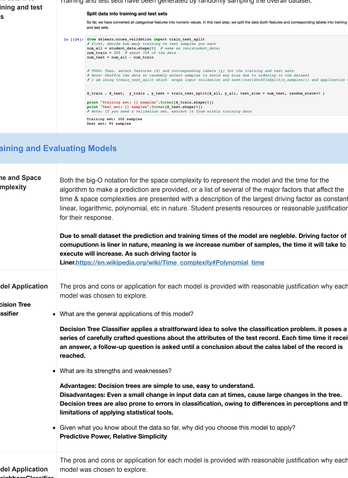
Preparing the Data

Identify feature and target columns

Code has been executed in the iPython notebook, with proper output and no errors.

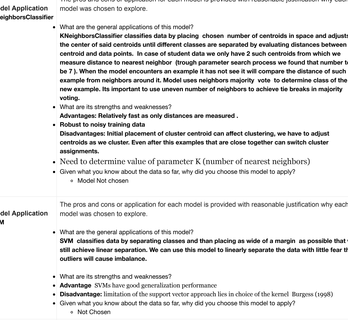
Preprocess feature columns

Code has been executed in the iPython notebook, with proper output and no errors.



Split data into training and test sets

Training and test sets have been generated by randomly sampling the overall dataset.



Training and Evaluating Models

Time and Space Complexity

Both the big-O notation for the space complexity to represent the model and the time for the algorithm to make a prediction are provided, or a list of several of the major factors that affect the time & space complexities are presented with a description of the largest driving factor as constant, linear, logarithmic, polynomial, etc in nature. Student presents resources or reasonable justification for their response.

Due to small dataset the prediction and training times of the model are negleble. Driving factor of comupuations is liner in nature, meaning is we increase number of samples, the time it will take to execute will increase. As such driving factor is Liner.https://en.wikipedia.org/wiki/Time_complexity#Polynomial_time

Model Application

The pros and cons or application for each model is provided with reasonable justification why each model was chosen to explore.

Decision Tree Classifier

- What are the general applications of this model?

  Decision Tree Classifier *applies a straitforward idea to solve the classification problem. It poses a series of carefully crafted questions about the attributes of the test record. Each time time it receive an answer, a follow-up question is asked until a conclusion about the calss label of the record is reached.*

- What are its strengths and weaknesses?

  Advantages: Decision trees are simple to use, easy to understand.
  Disadvantages: Even a small change in input data can at times, cause large changes in the tree. Decision trees are also prone to errors in classification, owing to differences in perceptions and the limitations of applying statistical tools.

- Given what you know about the data so far, why did you choose this model to apply?
  Predictive Power, Relative Simplicity

Model Application KNeighborsClassifier

The pros and cons or application for each model is provided with reasonable justification why each model was chosen to explore.

- What are the general applications of this model?
  KNeighborsClassifier classifies data by placing chosen number of centroids in space and adjusts the center of said centroids until different classes are separated by evaluating distances between the centroid and data points. In case of student data we only have 2 such centroids from which we measure distance to nearest neighbor (trough parameter search process we found that number to be 7). When the model encounters an example it has not see it will compare the distance of such example from neighbors around it. Model uses neighbors majority vote to determine class of the new example. Its important to use uneven number of neighbors to achieve tie breaks in majority voting.
- What are its strengths and weaknesses?
  Advantages: Relatively fast as only distances are measured .
- **Robust to noisy training data**
  Disadvantages: Initial placement of cluster centroid can affect clustering, we have to adjust centroids as we cluster. Even after this examples that are close together can switch cluster assignments.
- Need to determine value of parameter K (number of nearest neighbors)
- Given what you know about the data so far, why did you choose this model to apply?
  - Model Not chosen

Model Application SVM

The pros and cons or application for each model is provided with reasonable justification why each model was chosen to explore.

- What are the general applications of this model?
  SVM classifies data by separating classes and than placing as wide of a margin as possible that will still achieve linear separation. We can use this model to linearly separate the data with little fear that outliers will cause imbalance.

- What are its strengths and weaknesses?
- **Advantage** SVMs have good generalization performance
- **Disadvantage:** limitation of the support vector approach lies in choice of the kernel Burgess (1998)
- Given what you know about the data so far, why did you choose this model to apply?
  - Not Chosen

Model Performance Metrics

All the required time and F1 scores for each model and training set sizes are provided within the chart given. The performance metrics are reasonable relative to other models measured.

Student Intervention System – Decision Tree Classifier

	Training set size		
	100	200	300
Training time (secs)	0	0	0
Prediction time (secs)	0	0	0
F1 score for training set	0.719	0.810	0.807
F1 score for test set	0.769	0.758	0.758

Student Intervention System – LinearSVC Classifier

	Training set size		
	100	200	300
Training time (secs)	0.007	0.018	0.024
Prediction time (secs)	0	0	0
F1 score for training set	0.577	0.594	0.809
F1 score for test set	0.686	0.686	0.686

Student Intervention System – KNeighborsClassifier Classifier

	Training set size		
	100	200	300
Training time (secs)	0.001	0.001	0.001
Prediction time (secs)	0.002	0.002	0.008
F1 score for training set	0.774	0.760	0.806
F1 score for test set	0.714	0.714	0.714

Choosing the Best Model

Choosing the Optimal Model

Justification is provided for which model seems to be the best to use given computational cost and model accuracy.

Based on tests performed, I compared 3 models, Decision tree clasifier, Random forest, and Support Vector Classification. After evaluating performance of each model I choose Decision Tree Classifier Model for number of reasons:

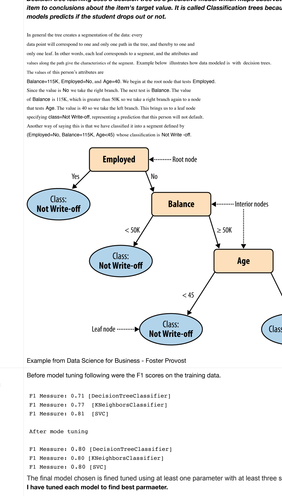
- Accuracy: Model predicts with high F1 Score. F1 score convys the balance between the precision and the recall. Comparing 3 models on the test set Decision tree classifier is able to obtain 75% accuracy which is 5% higher than the other models.

- Peformance: Obtainings training and test prediction fastes

Describing the Model in Layman's Terms

Student is able to clearly and concisely describe how the optimal model works in laymen terms to someone what is not familiar with machine learning nor has a technical background.
Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is called Classification trees because tree models predicts if the student drops out or not.

In general the tree creates a segmentation of the data every
data point will correspond to one and only one path in the tree, and thereby to one and
only one leaf. In other words, each leaf corresponds to a segment, and the attributes and
values along the path give the characteristics of the segment. Example below illustrates how data modeled is with decision trees.
The values of this person's attributes are:
Balance=115K, Employed=No, and Age=40. We begin at the root node that tests Employed.
Since the value is No we take the right branch. The next test is Balance. The value
of Balance is 115K, which is greater than 50K so we take a right branch again to a node
that tests Age. The value is 40 so we take the left branch. This brings us to a leaf node
specifying class=Not Write-off, representing a prediction that this person will not default.
Another way of saying this is that we have classified it into a segment defined by
(Employed=No, Balance>=115K, Age<45) whose classification is Not Write -off.



Model Tuning

Example from Data Science for Business - Foster Provost

Before model tuning following were the F1 scores on the training data.

```
F1 Measure: 0.71 [DecisionTreeClassifier]
F1 Measure: 0.77  [KNeighborsClassifier]
F1 Measure: 0.81  [SVC]

After mode tuning

F1 Measure: 0.80 [DecisionTreeClassifier]
F1 Measure: 0.80 [KNeighborsClassifier]
F1 Measure: 0.80 [SVC]
```

The final model chosen is fined tuned using at least one parameter with at least three settings.
I have tuned each model to find best parmameter.

Once the models have been tuned following are f1 scorces on test data:

```
F1 Measure: 0.76 [DecisionTreeClassifier]
F1 Measure: 0.68 [KNeighborsClassifier]
F1 Measure: 0.71 [SVC]
```

I have used grid search to search for the best parameters to use in the model. Grid search was done on all available data as only the parameters are being chosen, vs model built.

Tuned F1 Score

The F1 score is provided from the tuned model and performs better than the default model chosen.
Best F1 score for test set is 0.76

Quality of Code

Functionality

Code reflects the description in the documentation.